中国SKA区域中心原型系统——软件平台

劳保强[1,2], 张迎康[1], 安涛[1*], 徐志骏[1], 郭绍光[1,3], 伍筱聪[1], 吕唯佳[1]

1. 中国科学院上海天文台, 上海 200030;
2. 云南大学物理与天文学院, 昆明 650500;
3. 中国科学院大学, 北京 100049
*联系人, E-mail: antao@shao.ac.cn

收稿日期: 2022-06-29; 接受日期: 2022-08-30; 网络出版日期: 2023-01-05
国家重点研发计划(编号: 2018YFA0404603)、中国科学院青年创新促进会项目(编号: 2021258)和国家自然科学基金(编号: 12041301, 11873079, U1831204)资助

摘要 平方公里阵列(Square Kilometre Array, SKA)射电望远镜将在多个科学方向取得革命性的突破, 而SKA软件系统是影响科学产品的关键因素之一. SKA区域中心是天文学家进行SKA数据分析、科学研究和学术交流的平台. 处理SKA科学数据的软件环境需要具备通用性、灵活性和高适应性. 中国科学家已经建成了中国SKA区域中心原型机, 部署了被大型超级计算机广泛使用的作业调度系统, 并安装了能够处理当前主流射电望远镜观测数据的天文软件, 还部署了多个科学数据处理管线, 以方便不同科学方向的观测数据的自动化并行处理. 本文介绍了中国SKA区域中心原型机的软件平台和处理SKA先导望远镜数据的管线, 包括低频连续谱成像管线、谱线成像管线以及甚长基线干涉测量数据处理管线. 国内外用户已经基于该平台成功开展了SKA相关科学研究. 该平台的建设和运行为未来全面建设中国SKA区域中心提供了宝贵的实践经验.

关键词 平方公里阵列, 区域中心, 软件平台, 科学数据处理管线

PACS: 95.55.Br, 07.05.Bx, 07.05.Hd, 95.85.Bh, 95.75.-z

1 引言

平方公里阵列(Square Kilometre Array, SKA)射电望远镜是最大的天文望远镜, 将为人类探索宇宙、解决共同关注的科学问题做出重大贡献[1]. 2021年7月1日, SKA第一阶段(SKA1)的建设正式启动, 预计于2029年底建成并投入观测[2]. SKA1占总建设规模的10%[3], 建成后, SKA1全规模运行每年将向科学用户提供约710 PB的科学数据[4]. 面对如此史无前例的数据量, 无论是天文领域还是计算机领域都面临着空前巨大的挑战[5].

为了向全球的SKA科学用户提供高质量的数据产品、基本的数据处理资源和便捷的科学服务, SKA国际组织正计划在主要成员国建立若干SKA区域中心(SKA Regional Centre, SRC)[6]. 这些SRC节点将组成SRC网络(SRC Network, SRCNet), SRCNet将是一个分布式科研平台, 为世界各地的天文学家提供天文大数据的处理与分析、建模和可视化等服务. SRC白皮

书定义了SRCNet的六大基本功能: 数据物流、基本通用功能、数据处理、数据归档和管理、可扩展的资源管理和分配以及用户支持[4]. 其中, 数据处理功能是提供计算与存储资源和相关的软件服务, 以及软件编程和执行环境, 同时能够分发或共享科学数据到各个SRCNet节点进行交互式分析.

作为SRCNet节点之一, 中国SKA团队于2019年完成了世界首台原型机即中国SKA区域中心原型机(China SRC-Prototype, CSRC-P)[7], 该原型机具备天文数据的高速国际网络传输、存储和批处理能力. CSRC-P实际上是一台功能齐全的小型超级计算机, 部署了超级计算机所需的软件系统, 如作业调度系统、文件系统、管理软件等. CSRC-P同时部署了常见的射电天文数据处理软件和工具(如, 通用天文软件应用程序包(Common Astronomy Software Applications, CASA)[8]、天文图像处理系统(Astronomical Image Processing System, AIPS)[9]等), 这些软件在各个科学方向被广泛使用. 此外, 为了满足不同科学方向的差异化数据和SKA大数据的处理需求, SKA团队建立了多个科学数据处理管线, 能够支持大规模并行数据处理, 如默奇森宽场阵列(Murchison Widefield Array, MWA)的银河系和银河系外全天(GaLactic and Extragalactic All-sky MWA, GLEAM)巡天数据处理管线[10,11]. 还开发了基于人工智能的数据处理软件[12–15], 例如, 基于深度学习的射电星系识别与分类软件("河图", HeTu)[14].

后面的章节将介绍CSRC-P的作业调度系统、软件平台和数据处理管线, 并介绍几个科学应用案例.

2 作业调度系统

CSRC-P的登录节点承载着密集的访问, 用户量很大, 所以用户被禁止直接在登录节点上运行软件程序. 用户的程序必须在具有更大容量和更多资源的计算节点上运行. 在没有部署任何资源管理工具的情况下, 计算节点对用户来说是一个黑盒子. 对于单节点作业或任务, 当没有安全问题且只有一个用户时, 可以直接从登录节点访问计算节点, 并使用(Secure Shell, SSH)远程访问命令运行作业; 对于多节点作业或任务, 需要在节点之间建立免密码登录, 并需要部署一个消息传递接口(Message Passing Interface, MPI)环境(如OpenMPI, MPICH). 这些都需要具备高水平的(High Performance Computing, HPC)专业知识, 对于超级计算机的初学者来说这种方式较为复杂且困难. 另外, 从多用户和长期运维的角度来看, 这种方式还存在以下不足: 用户的作业和计算资源不能统一管理, 导致用户使用过程中出现资源竞争或资源浪费; 资源的运行状态不能实时监控, 导致维护困难; 可扩展性差, 资源规模扩大后仍然需要对多节点任务重复进行一系列复杂的部署. 因此, 根据分布式集群或超级计算机模式, 为CSRC-P配备作业和资源管理工具是非常重要和必要的.

对于科学家来说, 使用超级计算机的主要目的是尽快获得实验结果. SKA[16]和SRC的超级计算机采用的是异构超级计算模式, 不同节点结构的多个集群通过高速网络连接[17–19], 共同运行不同的工作负载. SKA有很多科学方向[1,20], 每个方向都有不同的资源类型和资源需求, 因此CSRC-P采用了混合异构计算架构[7], 包括: (1) 用于传统HPC任务的23节点x86 CPU集群; (2) 用于人工智能(Artificial Intelligence, AI)任务的4节点GPU集群; (3) 用于计算密集型任务的10节点ARM CPU集群.

超级计算机作业的生命周期包括排队和等待时间、计算资源分配、作业初始化、使用分配的资源执行作业、结果保存和资源释放. SKA工作流程经常会涉及到几个不同规模的资源分配, 调度器根据估计的响应时间选择最佳资源分配. 在这种情况下, 资源分配不仅在规模上不同, 而且在硬件结构上也不同, 从而使调度复杂化. 用于作业和资源管理的工具被称为作业调度系统, 当前主流的作业调度系统有: Load Sharing Facility (LSF)[21], Sun Grid Engine (SGE)[22], Simple Linux Utility for Resource Management (SLURM)[23]和Open Portable Batch System (OpenPBS)[24], 它们的相关情况对比分析见表1所示. 从表1可以看出SLURM在各方面都有明显的优势, 而且是最具扩展性和开放性的开源程序. 新发布的SLURM版本具有高级功能, 如回填、公平共享、抢占、多优先级、提前预约等, 这些都是许多超级计算中心所关心的. 它不仅在许多Top500的超级计算机中使用, 包括中国的国家超级计算中心广州超算天河二号, 而且还广泛用于SKA先导望远镜的数据中心, 如用于澳大利亚平方公里阵列探路者(Australian Square Kilometre Array Pathfinder, ASKAP)科学数据存储与处

表 1　主流作业调度系统对比

Table 1　Comparison of mainstream job scheduling systems

名称	授权许可	支持平台	最大节点数	费用
LSF	商用	Linux, Windows	6000+	付费
SGE	开源	Linux, Windows	未知	免费
SLURM	开源	Linux	120000+	免费
OpenPBS	开源	Linux, Windows	50000+	免费

理的Pawsey [25]超级计算机、荷兰低频阵列(Low Frequency Array, LOFAR)射电望远镜的计算机集群 [26]和南非卡鲁阵列射电望远镜(Meer Karoo Array Telescope, MeerKAT)的计算机集群 [27]等.

综合上述考虑, CSRC-P也采用SLURM作为作业调度系统. CSRC-P的计算节点根据计算架构被分为三个队列: ARM计算节点(arm)、X86计算节点(all-x86-cpu)以及GPU计算节点(all-gpu). 前两个队列主要用于串行、多线程、多节点和分布式作业, 最后一个队列用于GPU加速作业或AI模型训练和测试作业. 目前部署的SLURM版本是18.08[1], 它支持提交和管理异构作业, 以满足各种科学数据处理程序或软件的执行要求. 关于作业提交的方法和例子, 参见CSRC-P github项目"Introductory-CSRC-P"[2], 或者查看SLURM官方使用手册[3].

为了进行严格统一的计算资源管理, 计算节点设计了可插入认证模块(Pluggable Authentication Modules, PAM)访问控制, 只允许普通用户在有作业运行的情况下以SSH方式登入计算节点. 利用SLURM, 用户已经正常开展了包括串行、多线程、多进程、分布式等多种科学数据处理任务, 表明该作业调度系统能够满足天文数据处理模式要求. 此外, 还在SLURM作业调度器中实现了作业调度的灵活性和对异质作业的支持. 评估实验显示, 根据不同队列之间的负载不平衡程度, 作业的响应时间平均为50 ms, 比不使用调度器的响应时间改善了30%.

当然, SLURM仍然有一些设计上的缺陷. 例如, SLURM虽然引入了对GPU的支持, 但是它的调度算法并没有针对GPU进行优化: 在SLURM中, 如果不指定节点的数量(使用-N选项), 而只指定CPU核心或任务的数量(使用-n)和每个节点的GPU数量(使用-gres=gpu:), 那么这实际上会导致同一作业的不同运行分配不同的GPU总量(因为分配的GPU总量取决于请求CPU核心所分配的节点数量). 我们将继续跟踪用户作业的执行情况, 并根据不同队列上的应用程序的性能测量来改进作业调度策略. 例如, 允许用户将时间和资源列表与模块列表一起作为一个sbatch选项来指定. 另外, 也将探索实施一个响应时间驱动的策略, 作为优先级方案的替代.

3　软件平台

SRCNet需要为来自世界各地不同SKA科学工作组用户提供数据处理和分析的软件平台, 因此CSRC-P软件平台需要为不同的SKA科学案例提供相应的软件编程和执行、并行, 以及提供更加灵活的容器存储和执行环境. 此外, 需要为每个科学应用案例部署数据处理与分析软件和工具, 并为每个科学案例开发数据处理管线(或工作流)和基于人工智能的数据处理方法. 目前, CSRC-P软件平台支持处理来自当代重要射电望远镜的科学数据, 如MWA, ASKAP, LOFAR, Very Large Array (VLA)和Very Long Baseline Interferometry (VLBI). 对于不同的科学观测, 软件平台可以支持中低频连续谱观测、偏振观测、谱线观测、脉冲星计时和搜索等观测数据的处理. 对于大规模的科学数据处理任务, 提供了各种MPI并行和GPU环境, 支持单节点/多节点并行处理任务和GPU加速任务以及人工智能任务. 平台环境既支持本地化软件环境, 也支持虚拟软件环境. 下面对这两种环境进行了详细介绍.

3.1　本地软件环境

本地软件环境主要是通过在本地系统进行编译安装获得的. 天文软件包括SKA的科学应用软件都是在不断发展和迭代的, 对编程工具和环境的版本会有不同程度的依赖性. 为了满足不同科学用户的软件环境需求, CSRC-P安装了不同版本的编译器、库和基础软件. 这些编译器或软件采用了在不同的编译设置下得到的可执行程序和链接库, 使用这些编译器或软件时,

1) https://github.com/SchedMD/slurm.
2) https://github.com/SHAO-SKA/Introductory-CSRC-P.
3) https://slurm.schedmd.com.

只需要对环境变量进行修改, 尽可能为用户提供便利. 但是, 由于在软件编译过程中调用了大量的第三方库以及软件之间存在的依赖关系, 在执行特定版本的软件时, 环境变量的修改会变得极其复杂.

为了让用户能够快速切换不同版本的编译器或软件, 进而开展不同的科学数据处理, CSRC-P采用超算通用的环境管理工具Environment Modules[28]来管理和维护本地软件环境. 用户可以利用该工具提供的avail, load, unload, swarp等命令, 进行可用软件的查看、加载、卸载和切换等. 更详细的使用方法可以查看该工具的帮助信息或者官网使用手册.

本地软件环境主要分为三部分: 编译器和编程库、天文数据处理软件, 天文数据分析工具软件, 分别详见表2–5.

表2列出了主要的编译器和编程库, 有: GNU编译套件gcc[29]、跨平台安装编译工具Cmake[30]、统一计算设备架构(Compute Unified Device Architecture, CUDA)工具包CUDA Toolkit[31]、Rust语言构建和包管理器Cargo[32]、MPI并行实现库OpenMPI和MPICH、Python编程环境包. CSRC-P为用户提供了不同版本号的编译器和编程库, 并根据用户的需求持续更新和升级. 大多数天文数据处理软件均可以用gcc和Cmake进行编译安装, 少数并行加速软件用MPI和CUDA进行编译安装, 极少数Rust语言软件需要Cargo进行编译.

CSRC-P的主要天文数据处理软件包见表3和4, 这些软件能够开展(Jansky Very Large Array, JVLA), MWA, ASKAP, LOFAR, VLBI等重要射电望远镜阵列的数据处理, 数据类型包括连续谱图像、脉冲星、谱线和电压获取系统(Voltage Capture System, VCS)等不同格式的数据. 这些天文数据处理软件包的主要软件和程序如下:

(1) 主要用于射电频率干扰(Radio Frequency Interference, RFI)标记与消减的软件Aoflagger[33]和Cotter[34], 其中Aoflagger既可以用于单口径射电望远镜, 也可以用于射电干涉阵列; Cotter是基于Aoflagger开发的针对MWA数据格式的专用RFI标记和消减软件.

(2) 用于射电数据校准的软件有: mwa-reduce, Sagecal[35], Prefactor[36], 其中mwa-reduce是MWA数据专用的校准软件, 目前代码未开源, 仅供MWA团队成员使用; Sagecal和Prefactor主要用于LOFAR数据的校准, Sagecal支持GPU和MPI并行加速, 因此在运行速度上有较大的优势.

(3) 大视场成像软件, 比如WSClean[37], 该软件集成了多种大视场成像算法、去卷积/洁化算法和成图技术, 例如w-stacking[37]、w-snapshot[38]、多尺度洁化[39]、各向同性非抽样小波变换(Isotropic Undecimated Wavelet Transform, IUWT)压缩感知[40]、图像域栅格化[41]等, 已经被广泛用于MWA和LOFAR数据处理.

(4) 用于校准与成像的其他软件(包)有: 实时系统(Real Time System, RTS)[42]、YandaSoft[43]、factor[44]和Difmap[45], 其中, RTS主要用于MWA偏振和再电离时期(Epoch of Reionization, EoR)数据的校准与成像, 支持MPI和GPU并行运行; YandaSoft主要用于ASKAP数据校准; factor主要用于LOFAR数据, 主要是解决方

表 2　CSRC-P上主要的编译器和编程库

Table 2　Main compilers and programming libraries on CSRC-P

名称	功能简介	主要依赖库/包
gcc	GNU编译器套件, 目前提供版本: 4.9.3, 5.3.0, 7.3.0, 8.3.0和9.3.0.	gmp, mpfr, mpc, isl
Cmake	一个跨平台的安装(编译)工具, 可以用简单的语句来描述所有平台的安装(编译过程), 目前提供的版本: 3.15.2和3.8.2.	gcc
CUDA Toolkit	CUDA工具包, 提供CUDA C语言和C++语言编译器、CUDA驱动以及相关工具和科学库. 目前提供8.0, 9.0, 10.0, 10.1, 11.1版本.	gcc
Cargo	是Rust的构建系统和包管理器, 用于构建代码、下载依赖库并编译这些库等.	gcc, git, curl, pkg-config, OpenSSL
OpenMPI	是一个开源的MPI实现库, 能够结合来自高性能计算社区的所有专业知识、技术和资源, 以构建可用的最佳MPI库.	g++
MPICH	是MPI标准的高性能和广泛可移植的实现.	gcc
Python	Python编程环境平台. 目前提供2.7, 3.6, 3.7和3.8版本.	gcc, openssl, zlib, libffi, tk

表 3 天文数据处理软件-I

Table 3 Softwares for astronomical data reduction-I

名称	功能简介	主要依赖库/包
Aoflagger	用于干涉仪或单口径射电望远镜数据的射电频率干扰标记/消减, 支持的干涉仪望远镜包括LOFAR, WSRT, VLA, GMRT, ATCA和MWA, 单口径望远镜包括Parkes和Arecibo 305 m.	casacore, fftw3, boost, libxml, lapack, cfitsio, gtkmm (可选), libpng
Cotter	MWA专用射电频率干扰消减、数据平均和数据格式转换软件.	erfa, libpal, AOFlagger, dysco
mwa-reduce	MWA数据处理软件集成包, 主要用于MWA数据标记、天空模型建立和校准等.	casacore, cfitsio, fftw3, gsl, boost, gsl
MWA-Tools	MWA工具包, 提供MWA观测数据接口、MWA tile beam和各种分析工具等.	Matploylib, pyephem, pyfits, pywcs, cfitsio, AIPY, psycopg2, scipy, postgresql-client
WSClean	快速大视场成像软件, 主要用于MWA, LOFAR等低频干涉阵列数据成像.	casacore, cfitso, fftw3, boost, gsl
Chgcentre	针对Measurement Set数据的相位中心修改软件.	casacore, cfitso, fftw3, boost, gsl
RTS	MWA实时系统, 主要用于MWA偏振和EoR数据校准和成像, 支持单/多CPU节点并行处理(MPI)和GPU加速处理.	mpich, cfitsio, fftw3, cblas, lapack, wcslib, hralpix, slalib, cuda
Miriad	射电干涉仪数据处理软件包, 主要用于澳大利亚致密阵(ATCA)的数据处理.	linpack, pgplot, rpfits, wcslib
ASKAPsoft	ASKAP数据处理软件, 该软件集成了RFI消减、校准、自校准、连续谱成像、源搜寻、谱线成像、图像拼接等方法. 目前提供0.24.0, 1.0.19和1.0.2版本.	Casacore, fftw3, cfitsio, LOFAR, mpich, boost, APLpy, apr, astropy, blas, gsl, healpix, Ice, lapack, log4cxx, matplotlib, pytz, pywcs, wcslib
YandaSoft	射电干涉数据校准与成像软件, 主要用于ASKAP数据.	lofar-common, lofar=blob, askap-askap, askap-imagemath, askap-scimath, askap-parallel, askap-accessors, log4cxx, casacore, gsl, boost, mpich
CASA	通用天文软件应用包, 主要用于ALMA和VLA等射电数据处理. 目前提供版本: 4.5.3, 4.6.0, 4.7.2, 5.0.0, 6.1.0.	自带
DSPSR	脉冲星天文时间序列的数字信号处理软件.	psrdata, cfitsio, cuda
PRESTO	脉冲星搜寻与分析软件.	fftw3, pgplot, tempo, glib, cfitsio
PSRCHIVE	一个用于分析脉冲星天文数据的开源C++开发库. 它实现了广泛的算法, 用于脉冲星计时、闪烁研究、极化校准、单脉冲工作、RFI抑制等.	pgplot, tempo2
SIGPROC	Pulsar信号处理软件.	cfitsio, pgplot, zlib, fftw3
Tempo	Pulsar timing数据分析软件包.	pgplot, cfitsio, fftw3
Tempo2	Pulsar timing软件包.	pgplot, cfitsio, fftw3

位依赖效应(Direction Dependent Effects, DDEs)的影响.

CSRC-P已经部署的通用射电天文数据处理集成软件包有: ASKAPsoft [46], CASA [8], Miriad [47], AIPS [9] 和Obit [48]. ASKAPsoft主要用于ASKAP数据处理, 适配HPC环境, 该软件包集成了RFI标记、数据校准、成像、射电源搜寻等算法代码, 还提供了ASKAP连续谱成像和谱线成像管线等供用户使用. CASA是阿塔卡玛大型毫米波/亚毫米波天线阵(Atacama Large Millimeter/submillimeter Array, ALMA)和JVLA的主要数据处理软件, 也可用于其他射电望远镜, 尤其是Casacore是SKA数据校准和处理的核心库. AIPS是一个用于支持射电干涉阵列观测数据处理和分析的软件包, 最初是为VLA而设计的, 它固有的通用性使其成为大多数射电干涉仪特别是VLA、美国甚长基线阵列(Very Long Baseline Array, VLBA)和欧洲VLBI网(European VLBI Network, EVN)的标准数据处理软件包. Miriad是为毫米波/亚毫米波多通道谱线数据处理设计的, 它是澳大利亚望远镜致密阵列(Australia Telescope Compact Array, ATCA)使用的射电干涉测量数据处理包, 可用于完成连续谱和谱线观

表 4　天文数据处理软件-II

Table 4　Softwares for astronomical data reduction-II

名称	功能简介	依赖库/包
vcstools	MWA VCS数据处理工具.	gcc, Cargo, pal, cfitsio, psrfits_utils, fttw3, xgpu, hyperbeam, cuda
Obit	Obit是一组用于处理射电天文数据的软件包, 特别是干涉测量和单口径OTF成像.	cfitsio, glib, fftw, zlib, boost, gsl, plplot, python
AIPS	天文图像处理系统, 主要用于VLA, MERLIN, GMRT, WSRT, ATCA等射电干涉仪的数据处理, 同时也可用于VLBI的数据处理.	perl, libx11, libxext, libxpm, libncurses5, libbsd, libedit
Difmap	射电干涉仪数据成图软件, 主要用于VLBI数据.	gcc, pgplot, X11
Sagecal	一个快速的分布式和GPU加速的射电天文数据校准软件, 主要用于LOFAR数据.	Cmake, MPICH, gcc, casacore, wcslib, cfitsio
Prefactor	是用于校正LOFAR (High Band Array, HBA)和LOFAR (Low Band Array, LBA)观测中的各种仪器和电离层效应的管线.	DPPP, LoSoTo, LSMTool, EveryBeam, RMextract, Python, AOFlagger, WSClean, IDG和APLpy
factor	是用于LOFAR数据方向依赖效应(Direction-Dependent Effects, DDEs)校准和生成低噪声与高分辨率大视场图像的管线工具.	WSClean, DP3, LSMTool, LoSoTo, jinja2, Shapely, APLpy, pyds9和Dysco

表 5　天文数据分析软件或工具

Table 5　Softwares or tools for astronomical data analysis

名称	功能简介	依赖库/包
wcstools	世界坐标系统(WCS)工具包	cfitsio, wcslib
Aegean	MWA图像数据专用搜寻软件, 也可以用于其他射电图像进行源搜寻(Python 包), 也集成了背景噪声评估工具(BANE)、多分辨率图像掩膜工具(MIMAS)等.	scipy, six, tqdm, numpy, astropy, healpy, lmfit
Sextractor	主要用于光学图像的源搜寻软件.	ATLAS, FFTw3
Duchamp	三维天文数据源搜寻软件, 主要用于射电谱线数据.	pgplot, cfitsio, wcslib
TOPCAT	星表分析工具.	Java
Dysco	射电干涉数据压缩软件.	casacore
Swarp	FITS图像重采样和拼接软件.	cfitsio, wcslib
Montage	天文图像拼接软件.	cfitsio, wcslib, healpix
SAOImageDS9	用于天文数据的图像显示和可视化工具.	automake, autoconf, X11, zlib, tk, tcl, xml2, Xft, xslt
CARTA	天文立方体分析和渲染工具, 是为ALMA, VLA和SKA探路者设计的下一代图像可视化和分析工具.	gcc, casacore, hdf5, blas, wcslib

测数据的一系列的处理流程. Obit是一组用于干涉测量和单口径射电望远镜的即时(On The Fly, OTF)成像的软件包.

还有一些其他有特定用途的软件和工具包, 如, WSClean的Chgcentre[4]工具主要用于修改Measurement Set格式数据的相位中心; MWA-Tools是MWA观测数据接口, 集成了MWA各种数据分析工具. 剩余的是用于脉冲星搜寻与计时的软件, 这些软件相对独立于上述以成像为主的软件包, 常用的有: DSPSR [49], PRESTO [50], PSRCHIVE [51], SIGPROC [52], Tempo [53], Tempo2 [54]和vcstools [55–57]. DSPSR是一种用于射电脉冲星的高性能、开源、面向对象的数字信号处理软件

4) https://wsclean.readthedocs.io/en/latest/chgcentre.html.

库和应用程序套件. PRESTO是一套大型脉冲星搜索和分析软件, 是当前脉冲星搜索的核心工具和软件, 该软件主要设计目的是从对球状星团的长时间积分观测中, 有效地搜索毫秒脉冲星[58], 目前发现的脉冲星大部分是由该软件处理与分析得到的. PSRCHIVE是一个用于分析脉冲星天文数据的开源C++开发库, 它实现了广泛的算法, 可用于脉冲星计时、闪烁研究、极化校准、单脉冲工作、RFI 抑制等. SIGPROC是一个软件包, 旨在标准化多种类型的快速采样脉冲星数据的初始分析. Tempo和Tempo2是用于脉冲星计时数据分析的程序. vcstools主要用于MWA脉冲星的VCS数据处理.

如表5所示, CSRC-P已经配置的天文数据分析软件或工具主要有: 世界坐标系统(World Coordinate Systems, WCS)工具包(WCSTools[59])、源搜寻软件(Aegean[60], Sextractor[61], Duchamp[62], HeTu[14])、星表分析工具(Tool for Operations on Catalogues And Tables (TOPCAT)[63])、数据压缩软件(Dysco[64])、图像拼接软件(Swarp[65], Montage[66])、图像可视化分析工具(SAOImageDS9[67], Cube Analysis and Rendering Tool for Astronomy (CARTA)[68]). 其中, Aegean, Sextractor和Duchamp是基于分量拟合的天体识别软件, 源搜寻软件Aegean主要用于MWA图像, Sextractor主要用于光学图像, Duchamp主要用于ASKAP三维谱线图像(谱线图像数据), 而HeTu (河图)是使用深度学习开发的、基于射电形态的天体识别和分类软件. TOPCAT是一个用于表格数据的交互式图形查看器和编辑器, 为天文学家提供分析和操作星表和其他表格所需的大部分功能, 支持输入多种天文中常用的文件格式数据(包括Flexible Image Transport System (FITS), Virtual Observatory Table (VOTable)和Common Data Format (CDF)), 并且可以添加更多格式, 该工具尤其擅长交互式匹配大型(数百万行)表格. Dysco主要用于压缩Measurement Set文件格式数据. Swarp和Montage均可以进行FITS图像的重组与拼接. SAOImageDS9和CARTA是对FITS图像进行可视化分析的工具软件, SAOImageDS9同时支持命令行方式和提供Python接口包. CARTA支持远程交互式图像分析且是专门为新一代射电望远镜的图像分析进行设计的, 在高维度大尺寸图像数据分析有一定的优势.

为了让用户能够在不同的处理器上使用, 上述所有软件均分别部署了x86和ARM版本.

3.2 虚拟环境

为了解决软件编译难、应用优化难、软件环境移植难等问题, 通常可以通过搭建虚拟机或容器等虚拟化环境来解决. 一方面, 虚拟环境下软件环境相对干净, 且用户具有超级用户root权限, 软件安装部署不需考虑复杂的软件版本和库的依赖关系, 比本地环境要简单. 另一方面, 虚拟机或容器可以进行打包, 并能够快速在新的机器上部署, 具有即插即用的效果. 由于容器只需要一个虚拟化的操作系统, 它更加适用于超级计算机. 因此, CSRC-P所有计算节点均部署了目前主流的两种容器环境: Docker[69]和Singularity[70]. 在与SLURM作业调度系统的结合方面, Docker存在以下缺陷: 调度管理器的资源限制无法施加到容器中; 多用户(非root)使用时产生的结果文件会存在访问权限问题; 在运行时产生了更多非必要的资源开销. 相比Docker, Singularity原生支持MPI和SLURM, 能够与SLURM无缝结合, 可以直接使用SLURM提交多节点作业, 且具有以下优势: 环境打包迁徙更容易, 没有复杂的缓存机制, 占用存储空间少; 没有守护进程, 用户在容器内外保持一致, 且不占用任务资源, 安全性更高. 因此, 推荐用户在CSRC-P上使用Singularity, 也会对使用Docker的用户进行技术支持. CSRC-P目前建立了MWA, ASKAP和LOFAR数据处理软件镜像文件, 供用户使用, 未来将把其中的开源软件镜像上传至DockerHub, 进行长期管理和维护. 此外, CSRC-P的运维团队会根据用户的需求, 协助用户编写镜像编译文件和部署镜像文件.

考虑到CSRC-P的网络安全, 计算节点通常禁止访问外网. 登录节点虽然能够访问外网, 但是该节点拉取的镜像的本地存储路径并不是共享的, 用户实际运行的计算节点无法访问. 为了确保用户能够在计算节点上使用公有仓库的Docker镜像, CSRC-P提出了一种解决方案, 如图1所示. 该方案的思路是: 用户先利用登录节点(x86或ARM)的网络, 从Docker公有仓库上拉取容器镜像, 然后将该镜像打包并推送到本地私有仓库, 最后在计算节点利用节点之间的内网拉取该容器镜像到计算节点的存储上. 其中, 本地私有仓库主要通过配

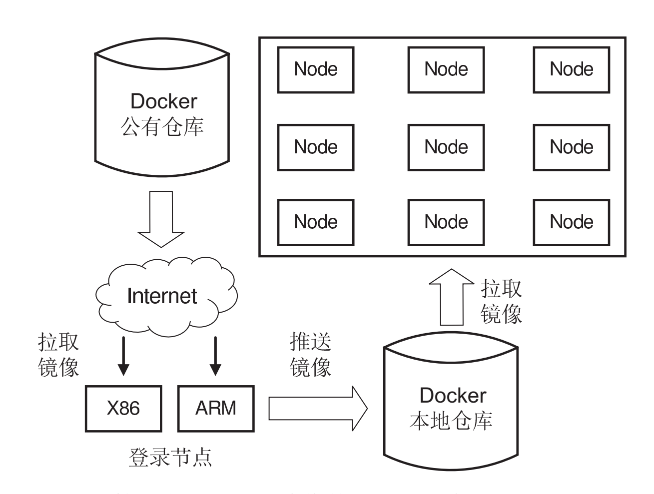


图 1 计算节点上Docker镜像使用解决方案

Figure 1 Docker image usage solution on compute node.

置登录节点5000端口并创建仓库服务获得.

4 数据处理管线及其应用案例

除了软件平台, CSRC-P还为科学用户提供基于软件平台的数据处理管线, 这些管线可以执行自动化和并行化的数据处理流程, 方便用户使用, 并已在实测数据上得到验证[10,11,71–73]. 科学用户通过一系列应用案例的实际操作, 能够快速掌握软件的使用方法和数据处理方法, 加快SKA的科学产出. 下面以3个典型科学应用案例为例, 介绍CSRC-P的数据处理管线系统.

4.1 谱线数据成像

基于中性氢(HI) 21 cm谱线的科学研究占了中国SKA十大科学方向中的3个: 宇宙再电离和宇宙黎明探测、中性氢巡天和宇宙学研究、中性氢星系动力学和星系演化[1], 足见其重要性. 21 cm谱线数据的主要处理流程包括: RFI标记、带通(Bandpass)校准、谱线成像和谱线源查找. 这里介绍的应用案例是谱线成像, 即输入的数据为已校准的数据.

测试数据来自ASKAP (36个12 m碟形天线)的(Deep Investigation of Neutral Gas Origins, DINGO) Pilot巡天[74]中GAMA 12天区的观测, 采用其中编号为0号的合成波束的观测数据, 观测起止时间为2019-3-12 14:03:22.1至2019-3-12 19:54:43.1, 共观测21081 s, 时间步长为10 s. 观测的相位中心为RA=177°3′, DEC=0°. 观测总频率通道数为15552, 起始频率为1151.5 MHz, 每个频率通道宽度为18.519 kHz, 总带宽为288 MHz, 中心频率为1295.4907 MHz.

本实验使用ASKAPsoft中的成像器imager[75]进行成像, 该成像器可以在分布式集群环境中运行, 也可以在独立的单机系统上运行, 数据分布灵活且内存占用少. 因此, 能够将测试数据按照频率通道划分, 进行多节点分布式处理, 从而提高处理速率. 本次实验使用了CSRC-P中的7个Intel x86 CPU节点, 共193个CPU核, 其中1个CPU核作为主进程, 不进行实际数据处理, 所以每个CPU核将处理15552/192=81个频率通道(nchanpercore)的数据. 整个谱线成像共消耗了11.7 h, 最终输出了一个大小为15552通道×2048像素×2048像素的立体图像FITS格式文件.

谱线成像主要的参数设置见表 6. ASKAP阵列的最大基线长度(MaxUV)为6 km, 因此MaxUV设置为6000 m. 观测角分辨率约为λ/MaxUV=3.52×10^{-5} rad (即约7.3 arcsec), 图像每个像素大小Image cellsize通常取角分辨率的四分之一, 本实验设置cellsize为2 arcsec. 观测视场(FoV)大小约为λ/D=0.0176 rad即1°, 输出图像大小约为FoV/Image cellsize=1800 像素, 且一般为2的正N次幂, 因此Image shape设置为2048×2048, 略大于视场大小. 成像算法选取W-projection[76,77], 因此数据栅格化gridder设为Wprojetion, w平面数(nwpanes)的取值由w最大值、输出图像像素大小和观测波长决定[78]. 为了节省内存, w平面数的取值不宜过大, 小于理论值也能获得较好的成像结果, 因此本实验的nwplanes设置为99.

为了进行结果分析, 利用软件平台中的Python包astropy[79]编写程序将输出的立方体图像分离为15552个二维图像FITS格式文件, 然后使用Aegean软

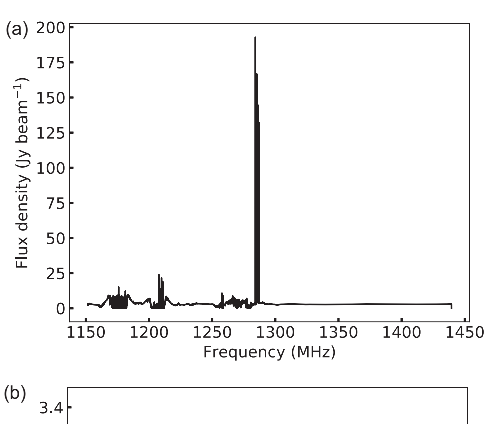
表 6 谱线成像主要参数设置

Table 6 Main parameter settings of spectral line imaging

参数	设置
MaxUV	6000 m
Images shape	[2048, 2048]
Images cellsize	[2, 2] arcsec
Image direction	[11h50m60.000, −00.26.59.96]
Images rest frequency	HI
nchanpercore	81
gridder	Wproject
Wproject nwplanes	99

件对分离后的图像进行源查找并输出每个图像的源表, 使用的是默认参数进行源查找. 通过分析每个源表, 获得了目标源的位置信息和峰值流量密度, 峰值流量密度随着频率的变化曲线如图2所示. 结合同一组数据的连续谱图像结果和图2的分析知, 在第8000通道之后为目标源信号, 之前的频率通道中均为RFI信号. 图2(b)在1.42 GHz位置附近没有明显的发射或吸收线出现. 该测试例子只是为了证明谱线成像流程已经正确部署, 得到的数据还需要进一步分析才能用于科学研究意义. 如图3为目标源信号频段中, 第10682通道的成像结果, 图中目标源图像的均方根(Root Mean Square, RMS)噪声与ASKAP相同宽度的单频率通道的RMS理论值相符[80], 进一步证明采用的成像流程输出的结果是正确的.

此外, 分别使用不同数量(19, 37, 73, 145和289个CPU核)的计算核进行了可拓展性实验(注: 每次实验均有1个CPU核作为主进程, 不用作观测数据处理), 消耗的时间分别是105.6, 47.1, 24.5, 13.6和10.1 h. 如图4所示, 是消耗时间和CPU核小时((CPU core)·h)随CPU核数增加的变化曲线. 从图4的消耗时间曲线可以看出, 随着CPU核数从19个增加到145个, 在双对数坐标系下, 消耗时间随着核数增加而准线性减少(线性坐标系下消耗时间则以近似幂律形式减少), 表明谱线成像的并行过程具有高度可扩展性. 另外, 从图4的CPU核小时曲线可以看出, 使用的CPU核数在19–145范围内, 计算效率较高; 当CPU核数为37时, 计算效率最高. 这种可扩展性实验对于未来SKA数据



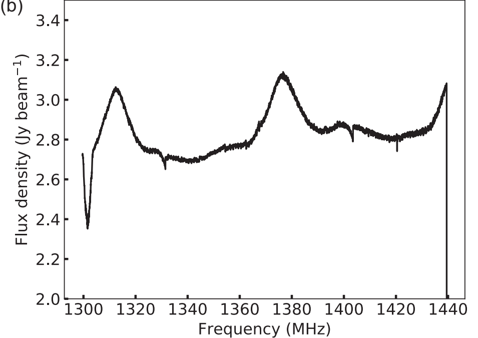


图 2 频率-峰值流量密度曲线. (a) 所有频率通道的结果; (b) 第8000个通道后的结果

Figure 2 Frequency-peak flux density. (a) The results of all frequency channels; (b) the results after the 8000th channel.

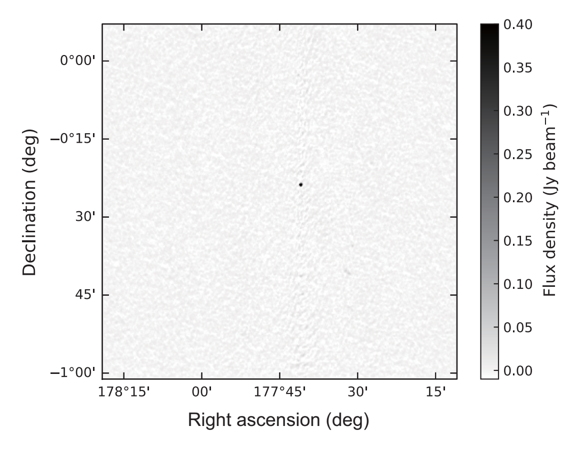


图 3 第10682频率通道成像结果. 目标源的峰值流量密度为2.7 Jy beam^{-1}, 图像的RMS噪声为3.1 mJy beam^{-1}

Figure 3 The imaging results of the 10682th frequency channel. The peak flux density of the target source is 2.7 Jy beam^{-1}, and the RMS noise of the image is 3.1 mJy beam^{-1}.

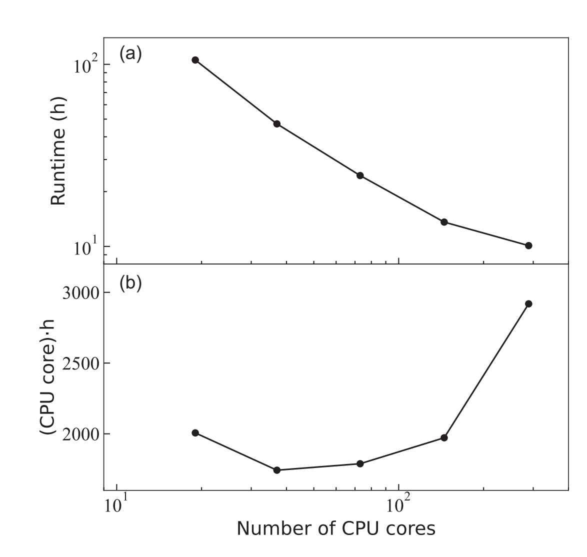


图 4 谱线成像运行时间(a)和CPU核小时(b)随CPU核数增加的变化曲线

Figure 4 Variation curves of spectral line imaging runtime (a) and CPU core hour (b) with increased CPU cores used.

处理有指导意义. SKA的谱线数据的频率通道数目达到65536个, 不仅是ASKAP的通道数目的4倍, 而且还要求精细通道成像, 生成的谱线立方体图像FITS文件的尺寸非常大. 因此, 有必要使用更多的计算核来外推检验“耗时-计算核”关系的有效性.

上述可拓展性实验的过程中, 除了耗时情况, 还记录了所有CPU核在整个数据处理过程中占用的最大内存量. 经过计算, 得到了每个CPU核平均占用的最大内存量约为3.1 GB, 并且使用不同CPU核数量, 每个CPU核平均占用的最大内存量基本相等. 按照当前使用的计算节点的总核数(32)计算, 每个计算节点占用的最大内存量约为99.2 GB, 表明在当前的CPU硬件配置下, 至少需要配置100 GB的内存(拓展实验中每个计算节点配置的内存为768 GB), 才能确保谱线成像管线正常处理ASKAP谱线数据. SKA的谱线巡天观测的角分辨是ASKAP的10倍以上, 在相同软件算法下, 单个计算节点将需要配置更大容量的内存.

4.2 连续谱成像管线

成像观测是除脉冲星和暂现源等时域科学方向以外的几乎所有科学方向都需要的基本模式. 本文介绍一个SKA低频阵列连续谱巡天数据处理管线的实例, 巡天观测获得的图像和星表等结果是开展SKA科学研究的基础数据. 该管线的主要处理步骤包括: 标记RFI、转换数据格式、建立天空模型、校准、深度成像、后处理(校准电离层引起的误差、修正射电源位置和流量密度)和图像拼接, 详细处理步骤介绍见文献[10, 81]. 在文献[10]的基础上, 本文进一步完善了建立天空模型、校准、深度成像和后处理步骤. 在天空模型的建立方面, 使用了更加完备的射电源模型制作该天区的天空模型[5), 能够弥补视场以外的空缺区域, 优化位于视场边缘的天体的校准. 在校准方面, 对观测数据视场内的亮源(例如Centaurus A[82]), 使用更加准确和完整的主波束模型(Full Embedded Element, FEE)进行主波束改正[83], 并进行自校准. 在深度成像方面, 使用了多尺度洁化方法[84]来提高延展源的成图质量. 在后处理方面, 采用了最新的程序(fits_warp[85])进行电离层改正. 通过这些改进, 获得的最终成图质量和天体信息的准确度更高.

由于连续谱巡天数据量大, 对每个数据进行顺序处理需要消耗大量的处理时间. 以MWA的GLEAM巡天为例, 共有高达6080个快照观测数据. 单个快照数据通常需要处理2–3 h, 如果使用单机不间断地依次处理快照数据, 需要700多天才能完成全部数据处理. 为此, CSRC团队开发了多节点分布式并行处理程序, 以实现大规模低频连续谱数据的自动化处理, 并分别部署在CSRC-P的x86 CPU节点和ARM CPU节点, 方便用户根据需求选择使用. 在文献[10]的基础上本文对并行策略做了改进, 提高了处理速度和可拓展能力. 在本文介绍的连续谱成像管线中, 每个节点只处理1个快照观测数据, 多个节点同时处理多个不同的数据. 基于改进后的管线, 能够一次性使用CSRC-P的所有CPU计算节点, 处理完成MWA GLEAM其中1个波段(200–231 MHz 波段, 共5个波段)总量约67 TB的6000多个快照数据, 仅需不到7天时间. 最终拼接获得的该波段全天区总强度图像的天区覆盖范围约为30939平方度, 平均RMS噪声约为8 mJy beam^{-1}, 与文献[81]发布的图像的RMS值一致. 该结果图像数据将用于活动星系核和宇宙磁场等研究. 图5展示了其中几个代表性天区的成像结果: 以活动星系核为主的射电点源、南天最亮最大结构的射电星系Centaurus A、银道面展源(用于超新星遗迹和宇宙磁场研究)、大麦哲伦星云、小麦哲伦星云和一个明亮的星系团. 从图5可以看出, 无论是致密的点源、低亮度的弥散源, 还是高亮度的大尺度延展源, 管线自动处理得到的图像质量较好.

使用Aegean软件进行射电源搜寻, 在4σ阈值以上检测到327621颗射电源(分量). MWA GLEAM团队发布了该巡天的河外星表[81]和银道面星表[86], 河外星表包含307455颗射电源或分量, 银道面星表包含22037射电分量. 使用CSRC-P软件平台中的Topcat软件将这两个星表合并, 获得了完整的GLEAM星表文件, 总射电分量数目为329492. 为了进行对比分析, 使用Topcat的Match Tables功能, 将本文200–231 MHz波段的星表结果与合并后的完整GLEAM星表进行交叉匹配, 匹配的最大偏差范围设置为(100 arcsec), 这是根据MWA GLEAM巡天的角

5) https://github.com/johnsmorgan/marco.

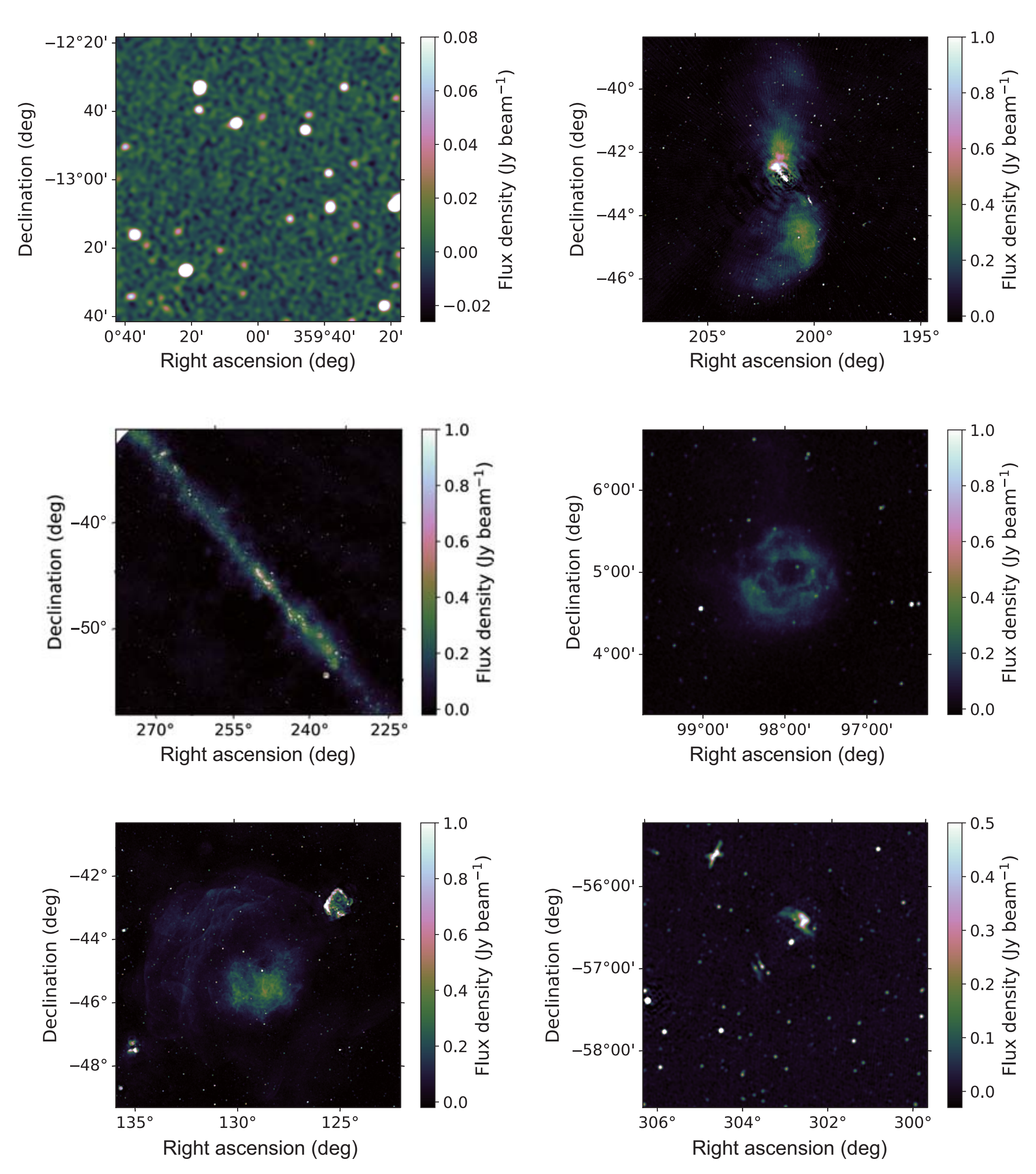


图 5 200–231 MHz波段全天区总强度图像中代表性区域的截图结果

Figure 5 The cutout result of the focal area in the total flux density image of the whole sky area on the 200–231 MHz band.

分辨率来定的[81]. 经过这一步, 获得了交叉匹配的结果, 共包含261236颗射电源/分量, 与公开发表的完整星表的匹配率为79.3%. 匹配星表的统计分析见图6, 图6(a)为匹配结果中RA之差($\Delta\alpha$)的柱状图, 图6(b)为匹配结果中DEC之差($\Delta\delta$)的柱状图, 两个图可以看出大部分匹配的射电源/分量的RA和DEC的偏差在50 arcsec (小于波束的1/2)范围以内, RA和DEC的占比分别为: 96.5%和95.6%. 这表明CSRC-P部署的管线获得的星表结果与公开发表的完整星表中匹配的射电源高度一致.

尚有20%左右没有匹配的射电源, 主要是由于本次仅处理了1个波段的数据, 有相当一部分射电源是陡谱, 它们在较低频率的其他四个波段被探测到, 在200–231 MHz波段的流量密度过低而没有被探测到. 本文获得的星表中多出的部分, 主要是来自已公开星表未包含的天空区域. 以上结果表明, CSRC-P的MWA低频

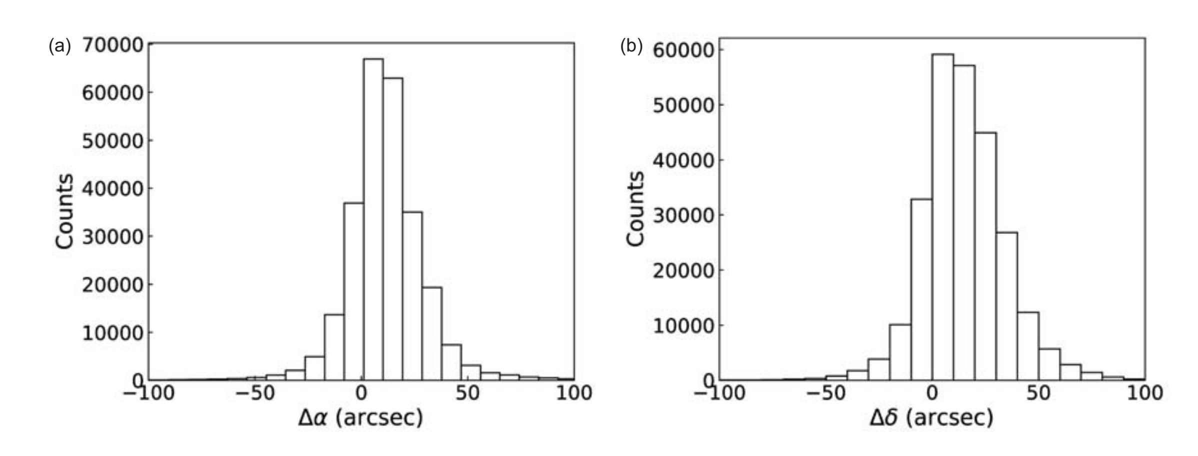


图 6 交叉匹配结果分析. (a) 匹配结果中RA之差($\Delta\alpha$)的柱状图; (b) 匹配结果中DEC之差($\Delta\delta$)的柱状图

Figure 6 Cross match result analysis. (a) A histogram of the difference of RA ($\Delta\alpha$) in the matching result; (b) a histogram of the difference of DEC ($\Delta\delta$) in the matching result.

连续谱成像管线获得的结果是可靠的, 图像可以直接用于天文研究.

4.3 VLBI管线

SKA1以核心密集阵为主, 基线长度为65–100 km, 类似于JVLA的联线干涉仪. 但是SKA第二阶段(SKA2)将是一个既包括核心密集阵又有延伸基线的扩展阵列, 最长基线达到3000–5000 km, 完整SKA2阵列的工作模式类似于VLBI模式. 实际上, SKA1有两条数据链路, 一条送到SKA的相关处理器, 另一条链路将原始数据直接送到VLBI的数据处理中心与其他VLBI望远镜观测的数据一起分析. 因此, VLBI也是SRC的数据处理模式之一. 在近期的SKA学术研讨会上6), 天文学家重点讨论了SRC对SKA-VLBI的支持.

相比于JVLA这样的联线干涉仪, VLBI观测和数据处理的复杂度较高, VLBI数据一直有处理时间长、处理难度大、人工介入需求多等特点, 且不同VLBI网和不同观测项目对于数据处理的要求也不尽相同, 因此国际上基本没有统一的VLBI数据处理管线. 一些大型巡天项目组往往自主开发特定管线, 只满足团队内项目的需求. 对于独立的个人观测项目, 需要科学用户自行处理, 而由于VLBI数据处理流程的复杂性, 用户的学习成本较高, 导致掌握VLBI数据处理的用户相对较少, 限制了VLBI的用户群体的规模.

CSRC团队考虑到用户对于VLBI数据处理有需求但是开发VLBI管线的时间成本很高的难题, 基于团队长期从事VLBI观测研究的丰富经验, 开发了一套VLBI数据处理管线, 能够满足绝大多数VLBI观测模式的数据处理. CSRC-P软件平台的建设目标也包含了面向SKA1-VLBI和SKA2对VLBI管线的前瞻性开发需求, 将开发出适合多种应用场景且处理多类型VLBI数据的管线, 从而更好地支持科学用户开展VLBI天体物理研究. 目前已经完成传统VLBI数据处理管线开发, 所涉及数据处理软件已经在CSRC-P软件平台中安装, 主要有AIPS和Difmap. 由于AIPS所用编程语言为Fortran, 且依赖交互式界面, 不利于进行大规模数据批处理, 因此管线主要采用AIPS提供的Python接口包ParselTongue [87]进行基于Python语言的开发. 数据处理的流程主要参考(National Radio Astronomy Observatory, NRAO)发布的AIPS COOKBOOK7).

VLBI数据处理管线的基本流程如下:

(1) 数据读入: 该步骤主要利用AIPS命令fitld读取可见度(Visibility)数据.

(2) 数据检查: 该步骤运行了一系列数据检查的AIPS命令, 主要包括listr, snplt和possm.

(3) 电离层改正和地球自转参数(Earth Orientation

6) 面向SKA时代的VLBI科学研讨会. https://whova.com/web/vlbis_202111/.
7) http://www.aips.nrao.edu/cook.html.

Parameters, EOP)改正: 首先从数据中提取观测当天的日期信息(Day of Year, DOY), 进而通过一个下载脚本从NASA空间测地数据中心[8)]下载观测对应当天的卫星测地数据(电离层模型以及地球定向参数), 并通过AIPS命令tecor和clcor将改正结果应用到校准表(CL table)中进行迭代.

(4) 幅度改正: 利用AIPS命令apcal来进行可见度数据的幅度改正, 并应用台站的天气信息与大气的不透明度估算, 之后将改正因子利用clcal迭代到校准表中.

(5) 星位角改正: 通过AIPS命令clcor的PANG项改正星位角引起的相位误差, 并将结果迭代到校准表.

(6) 手动相位校准: 通过手动选取条纹搜索源的一段观测时间(Scan), 使用fring程序进行条纹拟合, 利用所得结果来改正天线仪器的相位误差, 并利用clcal将结果迭代到新一个校准表中.

(7) 条纹拟合: 再次使用fring程序, 对所有的校准源在所有观测时间进行全局条纹拟合, 拟合采用的解间隔(Solint)由于在不同的观测频率有不同的优选值, 因此solint可以手动设置.

(8) 带通校准: 利用AIPS命令bpass来校准不同中频(IF)间的相位(和幅度)跳变, 主要运用了对主校准源(亮源)的互相关数据来作为校准依据. 结果将产生一个带通校准表(BP Table), 用于后续的校准和数据导出.

(9) 结果检查: 在上述校准完成后, 将通过检查命令snplt, possm来对校准后的相位, 幅度等信息进行检查, 若校准结果可接受则可进行下一步的数据导出.

(10) 数据导出: 该步利用AIPS命令split将迭代后的校准表(CL和BP)的校准信息应用到目标源中, 并导出为单源的可见度数据, 用于进行后续的成图等操作.

(11) 数据成图: 将校准好的可见度数据导入到Difmap软件包中进行自校准和成图操作.

图7和8展示的是利用VLBI管线对最近的一次4.6 GHz VLBA观测数据的处理结果. 目标源为VIK2318, 是一颗z=6.44的高红移射电类星体[88], 采用相位参考模式观测, 校准源是类星体J2314-3138. 校准前后相位随频率通道的变化见图7, 应用校准后, 相位趋于0°附近, 幅度的范围接近源的真实流量密度值. 图8展示的是校准源(a)和目标源(b)的成图结果图像, 其中校准源的RMS噪声约为0.16 mJy beam^{-1}, 信噪比达到2500, 目标源的RMS噪声为39 μJy beam^{-1}. 图8中两幅图像的梯度图的最外圈为对应图像RMS噪声值的3倍, 对于校准源图像, 梯度增量为上一梯度的2倍, 对于目标源图像, 梯度增量为上一梯度的$\sqrt{2}$倍. 从图8可以看出, VLBI可见度数据经过VLBI管线校准后的图像结构清晰, 且目标源经过校准后的RMS噪声水平不超过理论噪声值的2倍, 这表明数据校准结果接近天线所能达到灵敏度理论值; 图像噪声水平和信噪比达到理论预期, 表明该管线可以提供可靠的科学数据结果. 该VLBI管线处理数据的另外一个优势是结果的可重复利用性, 便于比照检查和结果复现, 这得益于该管线的脚本化和极少的人工干预. 在近几年, 科学用户已经利用该管线完成了多项VLBI数据处理, 所产出的成果已发表在天文学学术期刊中[89–95].

5 总结

本文介绍了CSRC-P的作业调度系统、软件平台和射电天文数据处理管线. 作业调度系统采用了SLURM调度系统, 该系统被各大超级计算机广泛使用, 具有良好的可拓展性, 并且是开源的, 易于维护. 在软件平台的设计上, 重视通用性, 既提供本地软件环境也提供虚拟化环境, 本地环境使用Environment Modules工具进行软件环境管理, 用户能够利用简单的命令快速切换不同科学软件环境, 且易于升级与维护. 虚拟环境支持Docker和Singularity容器镜像环境, 且提供已编译的MWA, ASKAP和LOFAR等数据处理软件镜像方便用户使用. CSRC-P团队还开展了数据处理方法的优化和自动化并行管线的开发. 本文重点介绍了已构建的谱线成像管线、低频连续谱成像管线和VLBI数据处理管线, 并用实际观测数据为案例做了验证实验. 谱线成像管线还进行了可拓展性实验, 为规模化扩展提供技术参考. 实验表明构建的管线均能够成功处理数据并获得可靠的结果. CSRC-P已经具备了向国内外科学用户服务的能力, 并且正在产生相关的科学成果.

8) https://cddis.nasa.gov.

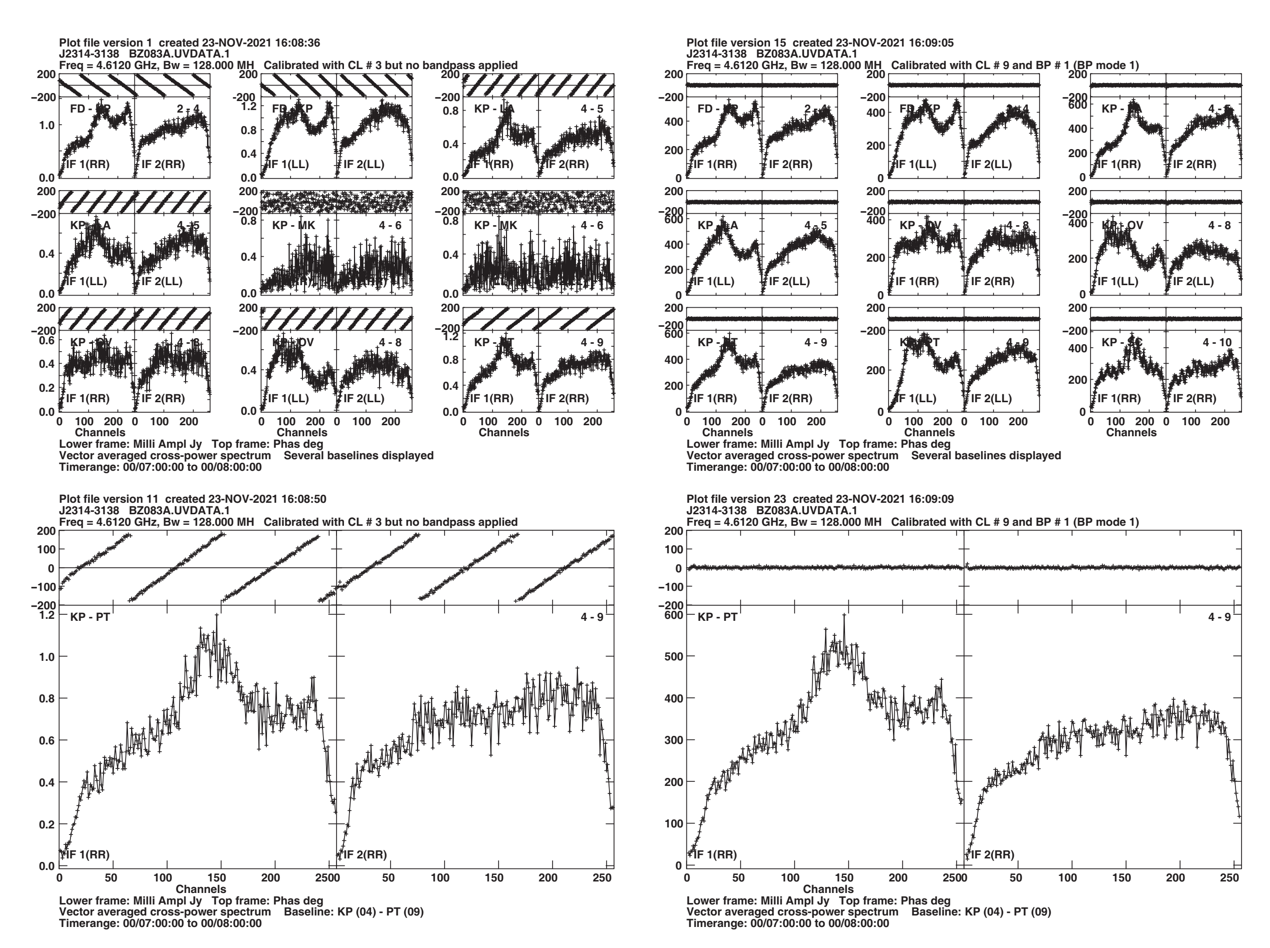


图 7 管线校准完成前后的参考源J2314-3138的相位-幅度对比图. 其中左边两图为校准前的相位-幅度图, 右边两图为校准后的相位-幅度图. 数据选取观测时间中某一小时时间段的观测进行积分, 以4号天线(KP)为参考天线. 下面两图为单一基线(KP-PT)右旋的相位-幅度图放大后的结果. 可以看出, 校准后数据的相位随channel的变化趋于0度, 幅度接近源的真实值

Figure 7 The phase-amplitude comparison diagram of the reference source J2314-3138 before and after the pipeline calibration is completed. The two pictures on the left are the phase-amplitude diagrams before calibration, and the two pictures on the right are the phase-amplitude diagrams after calibration. The data is selected for an hour during the observation time the observations in the period are integrated, and antenna No. 4 (KP) is used as the reference antenna. The following two pictures are the amplified results of the right-handed phase-amplitude diagram of a single baseline (KP-PT). It can be seen that in the phase of the data after calibration the change with the channel tends to 0 degrees, and the amplitude is close to the true value of the source.

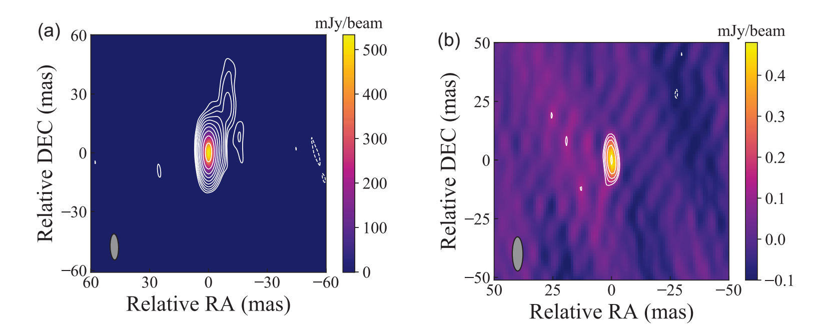


图 8 经过管线校准和成图处理获得的校准源J2314-3138和目标源VIK2318图像. (a) 校准源J2314-3138; (b) 目标源VIK2318 [96]

Figure 8 The imaging results of the calibrator source J2314-3138 and target source VIK2318 obtained by the calibration and imaging of the VLBI pipeline. (a) The image of the calibrator source J2314-3138; (b) the image of the target source VIK2318 [96].

致谢 感谢西澳大学Martin Meyer, Kevin Vinsen和Richard Dodson对谱线数据成像提供的帮助, 感谢科廷大学Natasha Hurley-Walker和Nick Seymour以及云南大学孙晓辉教授对连续谱成像提供的帮助. 本研究使用了中国SKA区域中心原型机的资源.

参考文献

1 Wu X P. Chinese SKA Science Book (in Chinese). Beijing: Science Press, 2019 [武向平. 中国SKA科学报告. 北京: 科学出版社, 2019]
2 Santander-Vela J, Bartolini M, Miccolis M, et al. From SKA to SKAO: Early progress in the SKAO construction. arXiv: 2110.13329
3 Garrett M A, Cordes J M, Deboer D R, et al. Square Kilometre Array: A concept design for phase 1. arXiv: 1008.2871
4 Quinn P, van Haarlem M, An T, et al. SKA Regional Centres. White Paper v1.0, 2020
5 An T. Science opportunities and challenges associated with SKA big data. Sci China-Phys Mech Astron, 2019, 62: 989531
6 An T, Wu X, Lao B, et al. Status and progress of China SKA Regional Centre prototype. Sci China-Phys Mech Astron, 2022, 65: 129501
7 An T, Wu X P, Hong X. SKA data take centre stage in China. Nat Astron, 2019, 3: 1030
8 McMullin J P, Waters B, Schiebel D, et al. CASA architecture and applications. In: Proceedings of Astronomical Data Analysis Software and Systems XVI ASP Conference Series. Tucson, 2007. 127–130
9 Greisen E W. Information Handling in Astronomy—Historical Vistas. Dordrecht: Springer, 2003
10 Lao B Q, An T. Deployment of SKA low frequency imaging system in China SKA Regional Centre (in Chinese). Sci Sin-Phys Mech Astron, 2020, 50: 059501 [劳保强, 安涛. SKA低频成像系统在中国SKA区域中心的部署. 中国科学: 物理学 力学 天文学, 2020, 50: 059501]
11 Wei J W, Zhang C F, Lao B Q, et al. Optimization of parallel processing of the Square Kilometre Array low-frequency imaging pipeline (in Chinese). Sci Sin-Phys Mech Astron, 2023, 53: 229503 [韦建文, 张晨飞, 劳保强, 等. SKA低频成像管线并行优化. 中国科学: 物理学 力学 天文学, 2023, 53: 229503]
12 Zhang H, Zhao Z, An T, et al. Pulsar candidate recognition with deep learning. Comput Electrical Eng, 2019, 73: 1–8
13 Liu X F, Lao B Q, An T, et al, Research on pulsar candidate identification method based on deep residual neural network (in Chinese). Acta Astronom Sin, 2021, 62: 20 [刘晓飞, 劳保强, 安涛, 等. 基于深层残差网络的脉冲星候选体分类方法研究. 天文学报, 2021, 62: 20]
14 Lao B, An T, Wang A, et al. Artificial intelligence for celestial object census: The latest technology meets the oldest science. Sci Bull, 2021, 66: 2145–2147
15 Xu Z J, An T, Guo S G, et al. A machine learning dataset for FRB detection in raw data (in Chinese). Sci Sin-Phys Mech Astron, 2023, 53: 229505 [徐志骏, 安涛, 郭绍光, 等. 一个面向原始数据搜寻的快速射电暴数据集. 中国科学: 物理学 力学 天文学, 2023, 53: 229505]
16 Farnes J S, Mort B, Dulwich F, et al. Building the world's largest radio telescope: The Square Kilometre Array science data processor. In: Proceedings of the 14th International Conference on e-Science. Amsterdam, 2018
17 Broekema P C, van Nieuwpoort R V, Bal H E. The Square Kilometre Array science data processor. Preliminary compute platform design. J Inst, 2015, 10: C07004
18 Scaife A M M, Joshi R, Cantwell T M, et al. Compute and storage for SKA Regional Centres. In: Proceedings of URSI Asia-Pacific Radio Science Conference (AP-RASC). New Delhi, 2019
19 Guo S G, An T, Xu Z J, et al. Progress and prospect of transcontinental high-speed data transmission at the Square Kilometre Array Regional Center in China (in Chinese). Sci Sin-Phys Mech Astron, 2023, 53: 229502 [郭绍光, 安涛, 徐志骏, 等. 中国SKA区域中心跨洲际高速数据传输进展及展望. 中国科学: 物理学 力学 天文学, 2023, 53: 229502]
20 Bourke T, Braun R, Fender R, et al. Advancing astrophysics with the Square Kilometre Array. In: Proceedings of Advancing Astrophysics with the Square Kilometre Array. Giardini Naxos, 2015
21 Natrajan A, Humphrey M A, Grimshaw A S. Grid Resource Management. Boston: Springer, 2004
22 Gentzsch W. Sun grid engine: Towards creating a compute power grid. In: Proceedings of the 1st IEEE/ACM International Symposium on Cluster Computing and the Grid. Brisbane, 2001. 35–36
23 Feitelson D, Rudolph L, Schwiegelshohn U. Job Scheduling Strategies for Parallel Processing. Berlin: Springer, 2003
24 OpenPBS—the Portable Batch System (PBS) Professional Open Source Project. https://www.openpbs.org
25 Kitaeff V V, Marrable D, Mararecki J T, et al. VO services with JPEG2000 client-server visualisation: Astronomy data services at Pawsey Supercomputing Centre. In: Proceedings of Astronomical Data Analysis Software and Systems XXV. Sydney, 2017. 69–72
26 Taffoni G, Becciani U, Bonafede A, et al. A distributed computing infrastructure for LOFAR Italian community. arXiv: 2201.11526
27 Heywood I. Oxkat: Semi-automated imaging of MeerKAT observations. Astrophysics Source Code Library, 2020
28 Furlani J L. Modules: Providing a flexible user environment. In: Proceedings of the 5th Large Installation Systems Administration Conference

(LISA V). San Diego, 1991. 141–152
29 Gough B J, Stallman R. An Introduction to GCC. Bristol: Network Theory Limited, 2004
30 Hoffman W, Martin K. The CMake build manager. Dr. Dobb's Journal: Software Tools for the Professional Programmer, 2003, 28: 40–43
31 Fatica M. CUDA toolkit and libraries. In: Proceedings of IEEE Hot Chips 20 Symposium. Stanford, 2008. 1–22
32 Klabnik S, Nichols C. The Rust Programming Language (Covers Rust 2018). San Francisco: No Starch Press, 2019
33 Offringa A R, van de Gronde J J, Roerdink J B T M. A morphological algorithm for improving radio-frequency interference detection. Astron Astrophys, 2012, 539: A95
34 Offringa A R, Wayth R B, Hurley-Walker N, et al. The low-frequency environment of the murchison widefield array: Radio-frequency interference analysis and mitigation. Publ Astron Soc Aust, 2015, 32: e008
35 Spreeuw J N, Yatawatta S, van Werkhoven B J C, et al. Scaling performance of the SAGECal calibration package: From LOFAR to SKA. In: Proceedings of the 33rd General Assembly and Scientific Symposium of the International Union of Radio Science. Rome, 2020
36 de Gasperin F, Dijkema T J, Drabent A, et al. Systematic effects in LOFAR data: A unified calibration strategy. Astron Astrophys, 2019, 622: A5
37 Offringa A R, McKinley B, Hurley-Walker N, et al. WSClean: An implementation of a fast, generic wide-field imager for radio astronomy. Mon Not R Astron Soc, 2014, 444: 606–619
38 Cornwell T J, Voronkov M A, Humphreys B. Wide field imaging for the Square Kilometre Array. In: Proceedings of the Image Reconstruction from Incomplete Data VII. San Diego, 2012
39 Rich J W, de Blok W J G, Cornwell T J, et al. Multi-scale clean: A comparison of its performance against classical clean on galaxies using things. Astron J, 2008, 136: 2897–2920
40 Starck J L, Bobin J. Astronomical data analysis and sparsity: From wavelets to compressed sensing. Proc IEEE, 2009, 98: 1021–1030
41 van der Tol S, Veenboer B, Offringa A R. Image domain gridding: A fast method for convolutional resampling of visibilities. Astron Astrophys, 2018, 616: A27
42 Mitchell D A, Greenhill L J, Wayth R B, et al. Real-time calibration of the murchison widefield array. IEEE J Sel Top Signal Process, 2008, 2: 707–717. arXiv: 0807.1912
43 Guzman J, Wicenec A. The rialto project: Software prototyping for the SKA science data processor based on australian precursor technologies. In: Proceedings of Astronomical Data Analysis Software and Systems XXIX. San Francisco, 2020. 531–534
44 Weeren R J, Williams W L, Hardcastle M J, et al. Lofar facet calibration. Astrophys J Suppl Ser, 2016, 223: 2
45 Shepherd M C. DIFMAP: An interactive program for synthesis imaging. In: Proceedings of Astronomical Data Analysis Software and Systems VI. San Francisco, 1997. 77–84
46 Guzman J, Whiting M, Voronkov M, et al. ASKAPsoft: ASKAP science data processor software. Astrophysics Source Code Library, 2019
47 Sault R J, Teuben P J, Wright M C H. A retrospective view of MIRIAD. In: Proceedings of the Astronomical Data Analysis Software and Systems IV. San Francisco, 1995. 433–436
48 Cotton W D. Obit: A development environment for astronomical algorithms. Publ Astron Soc Pac, 2008, 120: 439–448
49 van Straten W, Bailes M. DSPSR: Digital signal processing software for pulsar astronomy. Publ Astron Soc Aust, 2011, 28: 1–14
50 Ransom S. PRESTO: PulsaR exploration and search TOolkit. Astrophysics Source Code Library, 2011
51 Hotan A W, van Straten W, Manchester R N. PSRCHIVE and PSRFITS: An open approach to radio pulsar data storage and analysis. Publ Astron Soc Aust, 2004, 21: 302–309
52 Lorimer D R. SIGPROC: Pulsar signal processing programs. Astrophysics Source Code Library, 2011
53 Nice D, Demorest P, Stairs I, et al. Tempo: Pulsar timing data analysis. Astrophysics Source Code Library, 2015
54 Hobbs G B, Edwards R T, Manchester R N. Tempo2, a new pulsar-timing package—I. An overview. Mon Not R Astron Soc, 2006, 369: 655–672
55 Ord S M, Tremblay S E, McSweeney S J, et al. MWA tied-array processing I: Calibration and beamformation. Publ Astron Soc Aust, 2019, 36: e030
56 Xue M, Ord S M, Tremblay S E, et al. MWA tied-array processing II: Polarimetric verification and analysis of two bright southern pulsars. Publ Astron Soc Aust, 2019, 36: e025
57 McSweeney S J, Ord S M, Kaur D, et al. MWA tied-array processing III: Microsecond time resolution via a polyphase synthesis filter. Publ Astron Soc Aust, 2020, 37: e034
58 Manchester R N. Millisecond pulsars, their evolution and applications. J Astrophys Astron, 2017, 38: 42
59 Mink D J. WCSTools 3.0: More tools for image astrometry and catalog searching. In: Proceedings of the Astronomical Data Analysis Software and Systems XI. San Francisco, 2002. 169–172
60 Hancock P J, Trott C M, Hurley-Walker N. Source finding in the era of the SKA (Precursors): AEGEAN 2.0. Publ Astron Soc Aust, 2018, 35:

e011
61 Bertin E, Arnouts S. SExtractor: Software for source extraction. Astron Astrophys Suppl Ser, 1996, 117: 393–404
62 Whiting M T. DUCHAMP: A 3D source finder for spectral-line data. Mon Not R Astron Soc, 2012, 421: 3242–3256
63 Taylor M B. TOPCAT & STIL: Starlink table/VOTable processing software. In: Proceedings of the Astronomical Data Analysis Software and Systems XIV. San Francisco, 2005
64 Offringa A R. Compression of interferometric radio-astronomical data. Astron Astrophys, 2016, 595: A99
65 Bertin E, Mellier Y, Radovich M, et al. The TERAPIX pipeline. In: Proceedings of the Astronomical Data Analysis Software and Systems XI. San Francisco, 2002. 228–237
66 Berriman G B, Deelman E, Good J C, et al. Montage: A grid-enabled engine for delivering custom science-grade mosaics on demand. In: Proceedings of the Optimizing Scientific Return for Astronomy through Information Technologies. San Diego, 2004. 221–232
67 Joye W A, Mandel E. New features of SAOImage DS9. In: Proceedings of the Astronomical Data Analysis Software and Systems XII. San Francisco, 2003. 489–492
68 Wang K S, Comrie A, Harris P, et al. CARTA: Cube analysis and rendering tool for astronomy. In: Proceedings of the Astronomical Data Analysis Software and Systems XXIX. San Francisco, 2020. 213–216
69 Merkel D. Docker: Lightweight Linux containers for consistent development and deployment. Linux J, 2014, 2014: 2
70 Kurtzer G M, Sochat V, Bauer M W. Singularity: Scientific containers for mobility of compute. PLoS ONE, 2017, 12: e0177459
71 Guo S G, Lu Y, An T, et al. Scientific data flow and array simulation analysis for the SKA1 era (in Chinese). Sci Sin-Phys Mech Astron, 2023, 53: 229504 [郭绍光, 陆扬, 安涛, 等. 面向SKA1时代的科学数据流及阵列模拟分析. 中国科学: 物理学 力学 天文学, 2023, 53: 229504]
72 Wei J W, Zhang C F, Zhang Z L, et al. Parallel optimization of the pulsar search pipeline (in Chinese). Sci Sin-Phys Mech Astron, 2023, 53: 229506 [韦建文, 张晨飞, 张仲莉, 等. 射电脉冲星搜索的优化方法. 中国科学: 物理学 力学 天文学, 2023, 53: 229506]
73 Lao B, An T, Yu A, et al. Parallel implementation of w-projection wide-field imaging. Sci Bull, 2019, 64: 586–594
74 Meyer M. Exploring the HI universe with ASKAP. arXiv: 0912.2167
75 Wieringa M, Raja W, Ord S. ASKAPsoft pipeline gets ready for the pilot surveys. In: Proceedings of the Astronomical Data Analysis Software and Systems XXIX. San Francisco, 2020. 591–594
76 Cornwell T J, Golap K, Bhatnagar S. W projection: A new algorithm for wide field imaging with radio synthesis arrays. In: Proceedings of the Astronomical Data Analysis Software and Systems XIV. San Francisco, 2005. 86–90
77 Cornwell T J, Golap K, Bhatnagar S. The noncoplanar baselines effect in radio interferometry: The W-projection algorithm. IEEE J Sel Top Signal Process, 2008, 2: 647–657
78 Yu A, Lao B Q, Wang J Y, et al. Research on parallel algorithms for hybrid w-facets imaging (in Chinese). Progress Astron, 2020, 38: 421–435 [于昂, 劳保强, 王俊义, 等. 混合w-facets 成像并行算法研究. 天文学进展, 2020, 38: 421–435]
79 Robitaille T P, Tollerud E J, Greenfield P, et al. Astropy: A community Python package for astronomy. Astron Astrophys, 2013, 558: A33
80 Allison J R, Sadler E M, Bellstedt S, et al. FLASH early science—Discovery of an intervening HI 21-cm absorber from an ASKAP survey of the GAMA 23 field. Mon Not R Astron Soc, 2020, 494: 3627–3641
81 Hurley-Walker N, Callingham J R, Hancock P J, et al. GaLactic and extragalactic all-sky murchison widefield array (GLEAM) survey—I. A low-frequency extragalactic catalogue. Mon Not R Astron Soc, 2017, 464: 1146–1167
82 McKinley B, Tingay S J, Gaspari M, et al. Multi-scale feedback and feeding in the closest radio galaxy Centaurus A. Nat Astron, 2022, 6: 109–120
83 Sokolowski M, Colegate T, Sutinjo A T, et al. Calibration and stokes imaging with full embedded element primary beam model for the murchison widefield array. Publ Astron Soc Aust, 2017, 34: e062
84 Offringa A R, Smirnov O. An optimized algorithm for multiscale wideband deconvolution of radio astronomical images. Mon Not R Astron Soc, 2017, 471: 301–316
85 Hurley-Walker N, Hancock P J. De-distorting ionospheric effects in the image plane. Astron Comput, 2018, 25: 94–102
86 Hurley-Walker N, Hancock P J, Franzen T M O, et al. GaLactic and extragalactic all-sky murchison widefield array (GLEAM) survey II: Galactic plane $345° < l < 67°$, $180° < l < 240°$. Publ Astron Soc Aust, 2019, 36: e047
87 Kettenis M, van Langevelde H J, Reynolds C, et al. ParselTongue: AIPS talking Python. In: Proceedings of the Astronomical Data Analysis Software and Systems XV. San Francisco, 2006. 497–500
88 Ighina L, Belladitta S, Caccianiga A, et al. Radio detection of VIK J2318-3113, the most distant radio-loud quasar ($z = 6.44$). Astron Astrophys, 2021, 647: L11
89 Zhang Y K, An T, Frey S, et al. J2102+6015: A young radio source at $z = 4.575$. Mon Not R Astron Soc, 2021, 507: 3736–3744
90 Cheng X P, An T, Sohn B W, et al. Parsec-scale properties of eight Fanaroff-Riley type 0 radio galaxies. Mon Not R Astron Soc, 2021, 506:

1609–1622
91 An T, Mohan P, Zhang Y, et al. Evolving parsec-scale radio structure in the most distant blazar known. Nat Commun, 2020, 11: 1–8
92 Mohan P, An T, Yang J. The nearby luminous transient AT2018cow: A magnetar formed in a subrelativistically expanding nonjetted explosion. Astrophys J, 2020, 888: L24
93 Cheng X P, An T, Frey S, et al. Compact bright radio-loud AGNs. III. A large VLBA survey at 43 GHz. Astrophys J Suppl Ser, 2020, 247: 57
94 An T, Salafia O S, Zhang Y, et al. East Asia VLBI network observations of the TeV Gamma-ray burst 190114C. Sci Bull, 2020, 65: 267–271
95 Zhang Y K, An T, Frey S. Fast jet proper motion discovered in a blazar at $z = 4.72$. Sci Bull, 2020, 65: 525–530
96 Zhang Y K, An T, Wang A, et al. VLBI observations of VIK J2318-3113, a quasar at $z = 6.44$. Astron Astrophys, 2022, 662: L2

Software platform on China SKA Regional Centre prototype system

LAO BaoQiang[1,2], ZHANG YingKang[1], AN Tao[1*], XU ZhiJun[1], GUO ShaoGuang[1,3], WU XiaoCong[1] & LV WeiJia[1]

[1] *Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China;*
[2] *School of Physics and Astronomy, Yunnan University, Kunming 650500, China;*
[3] *University of Chinese Academy of Sciences, Beijing 100049, China*

The Square Kilometre Array (SKA) radio telescope is designed to revolutionize scientific breakthroughs in a variety of scientific fields, and the SKA software system is one of the key factors influencing scientific products. The software environment for processing SKA science data must be versatile, flexible, and adaptable. The SKA Regional Centre (SRC) serves as a platform for astronomers to analyze SKA data, conduct scientific research, and interact with other academics. To enable automated parallel processing of observational data from various scientific fields, Chinese scientists have developed the China SRC-prototype (CSRC-P), installed a job scheduling system widely used by large supercomputers, installed an astronomical software platform capable of processing observational data from current leading radio telescopes, and deployed multiple scientific data processing pipelines. This paper describes the software platform of the CSRC-P as well as pipelines for processing SKA precursor telescope data, such as the low-frequency continuum imaging pipeline, spectral line imaging pipeline, and very long baseline interferometry data processing pipeline. Users worldwide have successfully conducted scientific research on SKA using this platform. The knowledge gained from the construction and operation of this platform will be useful in constructing a full-scale SRC in the future.

Square Kilometre Array, reginal centre, software platform, scientific data processing pipeline

PACS: 95.55.Br, 07.05.Bx, 07.05.Hd, 95.85.Bh, 95.75.-z

doi: 10.1360/SSPMA-2022-0257